6. *Juuso Kaitila*, A content-based music recommender system, 2017, https://www.cs.rit.edu/usr/local/pub/GraduateProjects/2161/kxd8041/Report.pdf (Last accessed: 20.10.2023).

7. *Guan, Xin*, On reducing the data sparsity in collaborative filtering recommender systems., 2017, https://wrap.warwick.ac.uk/97978/1/WRAP_Theses_Guan_2017.pdf (Last accessed: 20.10.2023).

8. *Wanvimol Nadee*, Modelling user profiles for recommender systems, 2016, https://eprints.qut.edu.au/93723/1/Wanvimol_Nadee_Thesis.pdf (Last accessed: 20.10.2023).

9. *Andreu Vall, Matthias Dorfer, Hamid Eghbal-zadeh, Markus Schedl, Keki Burjorjee & Gerhard Widmer.* Feature-combination hybrid recommender systems for automated music playlist continuation, 2019, https://link.springer.com/article/10.1007/s11257-018-9215-8 (Last accessed: 10.10.2023).

10. *Sebastien Fremal, Fabian Lecron*. Weighting Strategies for a Recommender System Using Item Clustering Based on Genres, 2017, p. 6-11, https://www.sciencedirect.com/science/article/abs/pii/S0957417417300404 (Last accessed: 10.10.2023).

Надійшла до редколегії 13.10.2023 р

Ситнікова Поліна Едуардівна, канд. техн. наук, доцент, доцент кафедри системотехніки ХНУРЕ, м. Харків, Україна, e-mail: polina.sytnikova@nure.ua, ORCID: https://orcid.org/0000-0002-6688-4641.
Гребенюк Микита Олександрович, аспірант кафедри системотехніки ХНУРЕ, м Харків, Україна, e-mail: mykyta.hrebeniuk@nure.ua. ORCID: https://orcid.org/0009-0008-0989-7957.

УДК 004.627 DOI: 10.20837/0135-1710.2023.179.042

І.Г. ПЕРОВА, Н.С. МІРОШНИЧЕНКО

ОГЛЯД ІСНУЮЧИХ МЕТОДІВ ЗМЕНШЕННЯ РОЗМІРНОСТІ ТА КЛАСИФІКАЦІЇ ВЕЛИКИХ ВИБІРОК ДАНИХ

Аналіз великих вибірок даних, який проводиться з метою виявлення прихованих закономірностей і тенденцій, за останні роки стає все важливішим і кориснішим. Такі великі вибірки на поточний час характеризуються загальнодоступністю, складністю структур і великими розмірами.

Для вирішення проблеми великої розмірності даних пропонується ознайомлення з існуючими методами зменшення розмірності великих вибірок даних та порівняння ефективності цих методів на репозиторних вибірках. Розглядаються такі методи, як аналіз головних компонент (Principal Component Analysis), лінійний дискримінантний аналіз (Linear Discriminant Analysis), аналіз головних компонент ядра (Kernel Principal Component Analysis), багатовимірне масштабування (MDS), метод t-розподільного стохастичного вбудовування сусідів (t-SNE) та аналіз незалежних компонент (Independent Component Analysis). Як приклади великих вибірок даних використовуються набір даних ініціативи з нейровізуалізації хвороби Альцгеймера (ADNI) та набір даних про щитоподібну залозу, який є одним з декількох баз даних про щитоподібну залозу, доступних в репозиторії UCI.

1. Вступ

Для роботи з великими вибірками даних попередньо необхідно зменшити кількість параметрів вибірки. Цей процес називається зменшенням розмірності вибірки.

Зменшення розмірності як етап попередньої обробки машинного навчання є ефективним для видалення нерелевантних і надлишкових даних, підвищення точності навчання та покращення зрозумілості результату за допомогою візуалізації розмірності [1]. Дуже важливо зменшити розмірність набору даних без втрати будь-якої інформації, що міститься в них.

2. Мета і задачі дослідження

Метою даної роботи є порівняльний аналіз методів та підходів до зменшення розмірності великих вибірок даних у сфері машинного навчання.

Задачами дослідження є:

– ознайомлення з існуючими методами зменшення розмірності великих вибірок даних;

– опис головної ідеї кожного із методів;

– дослідження особливостей застосування кожного з методів на репозиторних вибірках даних;

– порівняння отриманих результатів.

3. Матеріали і методи дослідження

На даний момент найпопулярнішим методом для задач високої розмірності вважають метод аналізу головних компонент (Principal Component Analysis (PCA)) – метод вилучення ознак, який використовується для аналізу статистичних даних шляхом перетворення початкового набору даних у різноманітний набір лінійних комбінацій. Такі комбінації відомі як головні компоненти, що мають певні властивості щодо дисперсій. Виділення лінійних комбінацій робить розмірність системи щільнішою і в той же час зберігає інформацію про змінні зв'язки. Обчислення виконуються на наборі даних шляхом розрахунку власних значень коваріаційної матриці, розташованих у систематичному порядку за спаданням. Ця методика забезпечує максимальну реалістичність довільних рішень у просторі великої розмірності [2].

Як приклад розглянемо набір даних, який складається з переліку пацієнтів та ознак, що їх характеризують. Припустимо, що цей набір даних являє собою матрицю $X=\{x_{mn}\}$ розмірністю K×k, де K – кількість пацієнтів, а k – кількість ознак, які характеризують кожного пацієнта. Тоді кожен пацієнт описується (k×1)-вектором ознак

$$x(a) = (x_1(a), x_2(a), \dots x_j(a), \dots x_k(a))^T, \quad (1)$$

де a=1,…,K.

Перша головна компонента PC_1 є лінійною комбінацією вихідних змінних X_1, X2, ...,X_k

$$PC_1 = c_{11}X_1 + c_{12}X_2 + c_{13}X_3 \dots \dots . . c_{1p}X_k = \sum_{j-1}^{p} c_{ij}X_j \ . \quad (2)$$

Умова

$$c_{11}{}^2 + c_{12}{}^2 + c_{13}{}^2 \dots \dots . c_{1p}{}^2 = 1 \ , \quad (3)$$

вказує на те, що коефіцієнти c_{1p} мають бути нормованими, тобто квадрати суми їх коефіцієнтів дорівнюють 1. Ця умова гарантує, що власне значення PC_1 є найбільшим, враховуючи обмеження (3) на константу c. Обмеження має бути накладено для того, щоб уникнути збільшення власного значення PC_1 простим збільшенням одного або декількох значень вихідних змінних.

Проте, незважаючи на те, що PCA є дуже ефективним, він є і повністю неконтрольованою технікою зменшення розмірності. Це означає, що цей метод не використовує мітки класів під час навчання низьковимірного представлення, що призводить до субоптимальної продуктивності під час вирішення надалі задачі класифікації.

Ще одним з найважливіших фундаментальних методів є метод лінійного дискримінантного аналізу Фішера (Linear Discriminant Analysis (LDA)). Цей метод має дуже важливі властивості для класифікації даних. По-перше, він базується на простих геометричних принципах: необхідно максимізувати відстань між середніми значеннями класів (центрами) та мінімізувати внутрішньокласову дисперсію. Це означає, що LDA намагається знайти лінійну комбінацію ознак, яка найкращим чином розділяє класи, збільшуючи одночасно роздільну здатність. По-друге, метод LDA дозволяє зменшити розмірність даних, обираючи нові змінні (дискримінантні ознаки), які відображають основні відмінності між класами. Це корисно для подальшого аналізу та класифікації, оскільки він дозволяє працювати з меншими кількостями ознак, що може зменшити обчислювальні витрати і запобігти перенавчанню [3].

LDA зосереджується головним чином на проекції функцій з простору вищих вимірів у простір нижчих вимірів. Для цього необхідно обчислити відстань між середніми значеннями різних класів, яка називається міжкласовою дисперсією

$$S_B = \sum_{i=1}^{g} N_k (m_k - m)\,(m_k - m)^T \ , \tag{4}$$

де g – кількість класів; m – загальне середнє; m_k – середнє для відповідного класу; N_k – обсяг вибірки для k-го класу.

Наступним кроком розраховується внутрішньогокласова дисперсія (S_W)

$$S_w = \sum_{i=1}^{c} S_i \ , \tag{5}$$

де c – кількість класів; S_i – матриця розсіювання окремого i-го класу, яка обчислюється за формулою

$$S_i = \sum_{x \epsilon D_i}^{n_i} (x - m_i)\,(x - m_i)^T \ , \tag{6}$$

де D_i – множина прикладів з i-го класу; n_i – обсяг вибірки для i-го класу. m_i – середній вектор для i-го класу. Формула (6) показує, як внутрішньокласова дисперсія (S_W) обчислюється для кожного класу окремо. Матриця S_i – відображає розсіювання даних всередині конкретного класу i.

Останнім кроком необхідно побудувати простір нижчих розмірів, який максимізує дисперсію між класами та мінімізує дисперсію всередині класу.

$$P_{LDA} = \arg\max \left| \frac{P^T S_B P}{P^T S_W P} \right| \tag{7}$$

де P – критерій Фішера, або так звана проекція простору нижчої вимірності.

Проте даний метод має недоліки, основним з яких є неможливість надійної оцінки параметрів моделі без достатньої кількості зразків для кожного класу. Якщо деякі класи мають обмежену кількість прикладів, то LDA може давати неточні результати. Застосування LDA при малій кількості зразків в одному з класів схоже на спробу підігнати пряму до точки. Отже, без додаткових обмежень ці алгоритми будуть перенавчатися, тобто обиратимуть

класифікатор на основі шуму в даних, а не відкидатимуть шум на користь потрібного сигналу. Це означає, що даний метод можливо використовувати тільки для виконання контрольованого зменшення розмірності шляхом проекції вхідних даних у лінійний підпростір, що складається з напрямків, які максимізують поділ між класами. Розмірність результату зазвичай повинна бути меншою за кількість класів, тому це загалом досить сильне зменшення розмірності, яке має сенс лише у випадку багатокласовості [4].

Ще одним методом зменшення розмірності є аналіз головних компонент ядра (Kernel Principal Component Analysis (KPCA)). Стандартний PCA дозволяє зменшувати лише лінійну розмірність. Однак, якщо дані мають складнішу структуру, яка не може бути добре представлена в лінійному підпросторі, стандартний PCA не буде дуже ефективним. На щастя, ядро PCA дозволяє узагальнити стандартний PCA до нелінійного зменшення розмірності, крім того, даний метод дозволяє спрощувати обчислення. KPCA може обробляти багатовимірні набори даних із багатьма функціями, зменшуючи розмірність даних, зберігаючи найважливішу інформацію [5].

Порівняльний аналіз корисності функцій лінійного та нелінійного ядер PCA для розпізнавання за допомогою лінійного класифікатора виявив дві переваги нелінійного ядра PCA: по-перше, нелінійні головні компоненти мають кращі показники розпізнавання ніж відповідні показники лінійних головних компонент. По-друге, продуктивність для нелінійних компонентів можна додатково покращити, використовуючи більше компонентів, ніж це можливо в лінійному випадку [6].

Проте KPCA має і свої недоліки. Даний метод має деякі обмеження, наприклад, необхідність вибору відповідної функції ядра та її параметрів, що може бути складним і трудомістким. KPCA також може бути дорогим в обчислювальному плані для великих наборів даних, оскільки він вимагає обчислення матриці ядра для всіх пар точок даних.

Багатовимірне масштабування (Multidimensional Scaling (MDS)) – це ще одна методика зменшення та візуалізації, аналізу і звітності результатів. Даний метод на першому кроці працює з інформацією про певну форму несхожості між елементами набору об'єктів. Потім на основі цієї інформації створюється геометричне зображення, у якому відмінності представлені як відстані, які можуть бути використані для визначення ступеня схожості і відмінностей між парами об'єктів, кожна з яких представляє реальну точку даних. Об'єкти з невеликими відмінностями демонструють високий ступінь подібності, тоді як об'єкти з великими відмінностями вважаються несхожими [7].

Математичною основою MDS є функція напруження, яка вимірює різницю між відстанями у вихідному просторі та відстанями у просторі нижчої розмірності

$$stress = \sqrt{\frac{1}{2n^2 \sum_{i=1}^{n} \sum_{j=1}^{n} (d_{ij} - \widehat{d}_{ij})^2}}, \quad (8)$$

де dij – відстань між точками даних і та j у вихідному просторі; $\widehat{d_{ij}}$- відстань між точками даних і та j у просторі нижньої розмірності; n - кількість точок даних. Функція напруження є мірою відхилення відстаней у нижньовимірному просторі від відстаней у вихідному просторі і використовується для оцінки якості проекції.

Проте MDS є описовим методом, в якому практично повністю відсутнє поняття статистичного висновку [8].

До групи методів нелінійного зменшення розмірності можна також віднести метод t-розподіленого стохастичного вбудовування сусідів (t-SNE). Даний метод добре підходить для вбудовування даних великої розмірності з метою подальшої їх візуалізації в дво- або тривимірний простір низької розмірності. t-SNE знаходить закономірності в даних на основі схожості точок даних, схожість точок обчислюється як умовна ймовірність того, що точка A вибере точку B як свого сусіда. Умовна ймовірність для високої розмірності визначає, наскільки близько x_j знаходиться від x_i з урахуванням гауссівського розподілу навколо x_i із заданою дисперсією σ_i^2 , і розраховується за формулою:

$$p_{i|j} = \frac{\exp(-\frac{\|x_i - x_j\|^2}{2\sigma_i^2})}{\sum_{k \neq i} \exp(-\frac{\|x_i - x_j\|^2}{2\sigma_i^2})} , \quad (9)$$

Дисперсія σ_i^2 є різною для кожної точки; вона обрана таким чином, що точки в щільних областях мають меншу дисперсію, ніж точки в розріджених областях.

Для низькорозмірних аналогів y_i та y_j високорозмірних точок даних x_i і x_j можна обчислити подібну умовну ймовірність за формулою:

$$q_{j|i} = \frac{\exp(-\|y_i - y_j\|^2)}{\sum_{k \neq i} \exp(-\|y_i - y_k\|^2)} . \quad (10)$$

Вимірювання попарних відстаней у високо- та низьковимірному просторах з використанням різних розподілів ймовірностей дозволяє оптимальніше візуалізувати внутрішньокластерні деталі [9]. Метод є стохастичним, тому багаторазове його виконання з різними випадковими значеннями дасть різні результати. Допустимо запустити алгоритм кілька разів і вибрати вкладення з найменшою розбіжністю KL

$$KL = \sum_i \sum_{j \neq i} p_{ij} log \frac{p_{ij}}{q_{ij}} , \quad (11)$$

де pij та qij – парні ймовірності у високовимірному та низьковимірному просторі відповідно. Причиною точнішого визначення локальної структури є те, що функція витрат пропорційна парній ймовірності у високовимірному просторі. Відносно віддалені точки мають набагато менший вплив на функцію витрат.

Недоліком даного методу можна вважати великі кількість часу і обсяг простору, необхідні для обчислення. Це пов`язано з тим, що алгоритм обчислює попарно умовні ймовірності з метою мінімізувати суму різниць ймовірностей у вищих і нижчих вимірах, що потребує багато обчислень. t-SNE має квадратичну часову і просторову складність в залежності від кількості точок даних [10].

Наближеним до t-SNE є метод рівномірної багатовимірної апроксимації та проекції (UMAP), який використовується для візуалізації. Метод базується на трьох припущеннях:

– дані рівномірно розподілені на Рімановому просторові;

– Ріманова метрика є локально сталою (або її можна наближено вважати такою);

– простір є локально зв'язним.

Виходячи з цих припущень, можна змоделювати множину з нечіткою топологічною структурою. Вкладення знаходять шляхом пошуку низьковимірної проекції даних, яка має найближчу еквівалентну нечітку топологічну структуру.

Цей метод дуже добре працює з різними даними, від розпізнавання зображень до кластеризації та класифікації даних. Результати UMAP дуже якісні, та є конкурентоспроможними відносно методу t-SNE, завдяки своїм перевагам у швидкості та стабільності [11].

Аналіз незалежних компонент (Independent Component Analysis (ICA)) – це статистичний метод для виявлення прихованих факторів, які лежать в основі наборів випадкових величин, вимірювань або сигналів. ICA визначає генеративну модель для спостережуваних багатовимірних даних, які зазвичай подаються у вигляді великої бази даних зразків. У моделі передбачається, що змінні даних є лінійними сумішами деяких невідомих латентних змінних, і система змішування також невідома [12]. Приховані змінні вважаються негауссівськими та взаємно незалежними, і їх називають незалежними компонентами спостережуваних даних. Ці незалежні компоненти, які також називаються джерелами або факторами, можна знайти за допомогою ICA

$$x_i(t) = \sum_{j=1}^{m} a_{ij} s_j(t), \quad for\ all\ i = 1, \ldots, n\ , \tag{12}$$

де $x_i(t)$ – спостережувані змінні даних; i - індекс спостережуваної змінної даних, i=1,...,n; t – індекс часу або інший індекс різних спостережень, t=1,...,T. Припустимо, що $x_i(t)$ змодельовані як лінійні комбінації прихованих (латентних) змінних $s_j(t)$, j=1,...,m, з деякими невідомими коефіцієнтами a_{ij},

ICA поверхнево пов'язаний з PCA. Однак ICA є набагато потужнішою технікою, здатною знайти основні фактори або джерела, коли класичні методи повністю не спрацьовують. Дані, що аналізуються за допомогою ICA, можуть походити з різних сфер застосування, включаючи цифрові зображення, бази даних документів, економічні показники та психометричні вимірювання. У багатьох випадках вимірювання подаються у вигляді набору паралельних сигналів або часових рядів; для характеристики цієї задачі використовується термін "сліпе розділення джерел". Типовими прикладами є суміші одночасних мовних сигналів, які були вловлені кількома мікрофонами, мозкові хвилі, записані кількома датчиками, радіосигнали, що заважають, які надходять на мобільний телефон, або паралельні часові ряди, отримані в результаті якогось промислового процесу.

4. Результати дослідження та їх обговорення

Розглянуті вище методи було досліджено на наборі даних ініціативи з нейровізуалізації хвороби Альцгеймера (ADNI) [13]. Основною метою ADNI було перевірити, чи можна комбінувати для вимірювання прогресування легкого ступеня когнітивних порушень (MCI) і ранньої хвороби Альцгеймера (AD) результати магнітно-резонансної томографії (МРТ), позитронно-емісійної томографії (ПЕТ) та інших біологічних маркерів, а також клінічних та нейропсихологічних оцінок.

Набір даних містить записи про 90 ознак для 500 пацієнтів, які були відібрані випадковим чином (по 100 пацієнтів у кожному класі) для оцінки моделей. П'ять отриманих класів:

когнітивно нормальні (CN), хвороба Альцгеймера (AD), ранні легкі когнітивні порушення (EMCI), пізні легкі когнітивні порушення (LMCI) та значні порушення пам'яті (SMC).

На рис. 1 показано тривимірне відображення результатів застосування розглянутих методів зменшення розмірності. AD, CN, EMCI, LMCI та SMC позначені червоним, блакитним, зеленим, пурпуровим та помаранчевим кольорами відповідно.

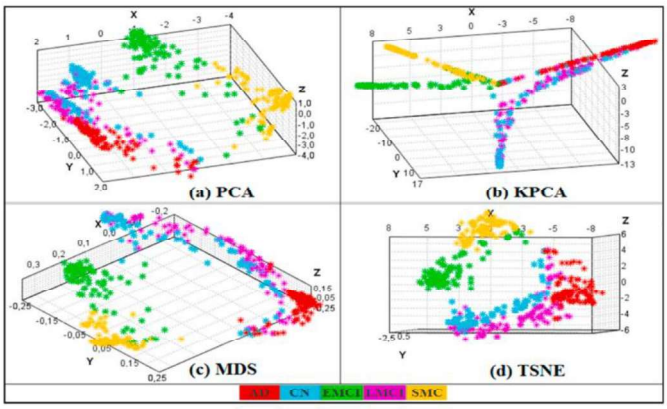

Рис. 1. 3D-візуалізація зменшень: a – PCA; b – KPCA; c – MDS; d – t-SNE.

Дослідження показали, що метод зменшення розмірності t-SNE показав відмінну класифікаційну оцінку даних. Даний метод є більш значущим і кращим для використання класифікаторами за інші методи.

Дослідження розглянутих методів було також проведено на наборі даних про захворювання щитовидної залози [14], який є одним з декількох баз даних про захворювання щитовидної залози, доступних в репозиторії UCI. У наборі даних зареєстровано 215 пацієнтів з трьох категорій: 1 – в нормі, 2 – страждають на гіпертиреоз, 3 – страждають на гіпотиреоз. Кожна вибірка містить п'ять характеристик: T3 реверсивний, тироксин, трийодтиронін, тиреотропний гормон, ТТГ.

	Normal	Suffer	Hyperthyroidism
Normal	136	5	9
Suffer	9	18	3
Hyperthyroidism	7	4	24

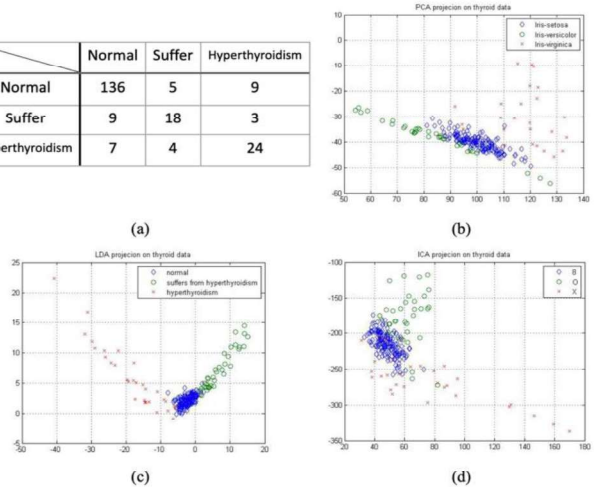

Рис. 2. Результати застосування методів зменшення розмірності для наборів даних щодо щитовидної залози: a – K-середніх; b – PCA; c –LDA; d – ICA

Рис. 2 ілюструє результати застосування методів PCA, LDA та ICA, отриманих для наборів даних про захворювання щитовидної залози.

Результати дослідження показують, що LDA має кращу продуктивність візуалізації, ніж PCA та ICA, оскільки використовує інформацію мітки. В алгоритмі ICA є випадковий фактор,

це означає, що якщо ми виберемо різні вектори для обчислення незалежних компонентів, продуктивність ICA може змінюватись.

5. Висновки

Ознайомившись із головними ідеями розглянутих у дослідженні методів, автори дійшли висновку, що кожен з цих методів має свої унікальні властивості і використовується в залежності від конкретних завдань та властивостей даних. PCA спрямований на зменшення розмірності шляхом проекції даних на нові ортогональні власні вектори (головні компоненти), які відображають напрямки найбільшої дисперсії в даних. LDA є методом зменшення розмірності, який спрямований на максимізацію віддільності класів даних шляхом проекції даних на нові відомі вектори (власні вектори Фішера). ICA намагається розкласти спостереження на лінійні комбінації незалежних компонентів, які максимально незалежні один від одного. KPCA є розширенням PCA, яке використовує ядерні функції для перетворення даних в вищу вимірність перед застосуванням PCA. Це дозволяє виявити нелінійні залежності в даних. t-SNE створений для візуалізації вимірюваних даних шляхом зменшення вимірності та збереження глобальних та локальних структур даних. Він базується на ймовірнісних розподілах схожості між об'єктами.

Проте, попри свою унікальність, ці методи мають спільні характеристики. Вони допомагають перетворити високорозмірні дані у низькорозмірний простір, зберігаючи при цьому важливу інформацію або структуру даних. Ці методи також є важливими інструментами в машинному навчанні та аналізі даних, оскільки вони допомагають зрозуміти та обробляти великі обсяги даних, зменшуючи їх розмірність або покращуючи їхню структуру.

Результати аналізу розглянутих методів дозволяють зробити висновок про відсутність на даний час досконалішого методу зменшення розмірності, який зміг би класифікувати багатокласові дані з мільйонами ознак, отримуючи при цьому сильні теоретичні гарантії, сприятливі та інтерпретовані емпіричні результати, а також гнучку, надійну та масштабовану реалізацію.

Перелік посилань:

[1] Ayesha, S., Hanif, M. K., & Talib, R. (2020, July). Overview and comparative study of dimensionality reduction techniques for high dimensional data. Information Fusion, 59, 44–58. https://doi.org/10.1016/j.inffus.2020.01.005.

[2] B. COY, “DIMENSION REDUCTION FOR ANALYSIS OF UNSTABLE PERIODIC ORBITS USING LOCALLY LINEAR EMBEDDING,” *International Journal of Bifurcation and Chaos*, vol. 22, no. 01, p. 1230001, Jan. 2019, doi: 10.1142/s0218127412300017.

[3] Badaoui, F., Amar, A., Ait Hassou, L., Zoglat, A., & Okou, C. G. (2017, October 10). Dimensionality reduction and class prediction algorithm with application to microarray Big Data. Journal of Big Data, 4(1). https://doi.org/10.1186/s40537-017-0093-4.

[4] S. Vannatta, “The Return of the Repressed (and Oppressed): A Freudo-Marxian Analysis of Jordan Peele’s Us,” *Popular Culture Review*, vol. 31, no. 2, 2020, doi: 10.18278/pcr.31.2.10.

[5] Essa, A. M., & Ghalib Alrawi, A. (2019, September 1). Comparison Between The Method of Principal Component Analysis And Principal Component Analysis Kernel For Imaging Dimensionality Reduction. IRAQI JOURNAL OF STATISTICAL SCIENCES, 16(29), 11–24. https://doi.org/10.33899/iqjoss.2019.164189.

[6] Jiang, J. L., Li, S. Y., Liao, M. L., & Jiang, Y. (2019). Application in Disease Classification based on KPCA-IBA-LSSVM. Procedia Computer Science, 154, 109–116. https://doi.org/10.1016/j.procs.2019.06.017.

[7] Dzemyda, G., Sabaliauskas, M., & Medvedev, V. (2022). Geometric MDS Performance for Large Data Dimensionality Reduction and Visualization. Informatica, 299–320. https://doi.org/10.15388/22-infor491.

[8] T. Li, Q. Yin, R. Song, M. Gao, and Y. Chen, “Multidimensional scaling method for prediction of lysine glycation sites,” *Computing*, vol. 101, no. 6, pp. 705–724, Mar. 2019, doi: 10.1007/s00607-019-00710-x.

[9] Spiwok, V., & Kříž, P. (2020, June 30). Time-Lagged t-Distributed Stochastic Neighbor Embedding (t-SNE) of Molecular Simulation Trajectories. Frontiers in Molecular Biosciences, 7. https://doi.org/10.3389/fmolb.2020.00132.

[10] Ayyappa. T and S. Kurse, "Fault Detection of Bearing using XGBoost Algorithm and Data Visualization using t-distributed stochastic neighbor embedding (t-SNE) Method," SSRN Electronic Journal, 2021, Published, doi: 10.2139/ssrn.3834976.

[11] J. Heaton, "Ian Goodfellow, Yoshua Bengio, and Aaron Courville: Deep learning," *Genetic Programming and Evolvable Machines*, vol. 19, no. 1–2, pp. 305–307, Oct. 2018, doi: 10.1007/s10710-017-9314-z.

[12] Hong, S. E. (2019, December 31). Exploring Independent Component Analysis Based on Ball Covariance. The Korean Data Analysis Society, 21(6), 2721–2735. https://doi.org/10.37727/jkdas.2019.21.6.2721.

[13] ADNI | Study Documents. (n.d.). https://adni.loni.usc.edu/methods/documents/;

[14] UCI Machine Learning Repository. (n.d.). https://archive.ics.uci.edu/dataset/102/thyroid+disease.

Надійшла до редколегії 02.08.2023

Перова Ірина Геннадіївна, доктор технічних наук, професор, професор кафедри системотехніки ХНУРЕ, м. Харків, Україна, e-mail: iryna.perova@nure.ua. ORCID: https://orcid.org/0000-0003-2089-5609.
Мірошниченко Неля Сергіївна, аспірант кафедри системотехніки ХНУРЕ, м. Харків, Україна, e-mail: nelia.miroshnychenko@nure.ua. ORCID: https://orcid.org/0000-0002-3846-1668.

УДК 004.89:61 DOI: 10.20837/0135-1710.2023.179.050

І.Ю. ПАНФЬОРОВА, А.С. БУЦЬКА

ДОСЛІДЖЕННЯ АНСАМБЛЮВАННЯ МОДЕЛЕЙ MACHINE LEARNING В МЕДИЧНІЙ ДІАГНОСТИЦІ

Розглянуто застосування моделей машинного навчання (machine learning – ML) в медичній діагностиці. Представлено основні виклики та цілі останніх досліджень у сфері медичного прогнозування. Основну увагу зосереджено на порівнянні існуючих моделей ML. Проведено аналіз вже існуючих рішень для розробки ансамблю моделей ML. Розраховано ключові характеристики моделей ML: точність, чутливість, специфічність та AUC-ROC. Запропоновано варіант об'єднання цих моделей в ансамбль для покращення точності – основної характеристики прогнозування діагнозу.

1. Вступ

Медична діагностика є критично важливим аспектом охорони здоров'я, оскільки її рівень безпосередньо впливає на точність результатів діагностування і, як наслідок, на вірність рішення щодо лікування пацієнтів. Нова ера в медичній діагностиці розпочалася з появою машинного навчання (machine learning – ML) , яке надало розширені інструменти для аналізу великих обсягів медичних даних, включаючи історії хвороби пацієнтів, зображення та генетичну інформацію. Використання моделей ML може сприяти виявленню прихованих закономірностей і аномалій, а також прогнозуванню ризиків розвитку того чи іншого захворювання, що робить їх цінними інструментом для медичних працівників.

Один з напрямів сучасних досліджень в галузі застосування ML у медичній діагностиці полягає у ліквідації недоліків існуючих моделей ML шляхом їх комбінації. Цей різновид ML, який забезпечує покращення продуктивності і точності прогнозів за рахунок комбінації кількох моделей ML, отримав назву «ансамблеве навчання» [1]. Мета досліджень за цим напрямом полягає у порівнянні існуючих моделей ML, пропозиції множини варіантів та виборі найкращого варіанту вирішення задачі створення ансамблю для прогнозування діагнозу.

Ансамблеве навчання, на жаль, не вільне від деяких недоліків. Так непростою задачею є інтеграція моделей ML, адже до того, як впровадити моделі ML у вже існуючі медичні